[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G06F 17/30

[12]发明专利申请公开说明书

[21]申请号 98804175.8

[43]公开日 2000年5月10日

[11]公开号 CN 1252876A

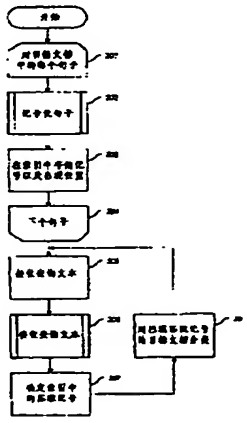
[22]申请日 1998.2.11 [21]申请号 98804175.8
[30]优先权
[32]1997.3.7 [33]US [31]08/886,814
[86]国际申请 PCT/US98/03005 1998.2.11
[87]国际公布 WO98/39714 英 1998.9.11
[85]进入国家阶段日期 1999.10.14
[71]申请人 微软公司
地址 美国华盛顿
[72]发明人 约翰·J·麦瑟利 乔治·E·海德恩
斯蒂芬·D·理查德森 威廉·B·杜兰
卡轮·杰森
[74]专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 鄭 迅

权利要求书4页 说明书17页 附图页数18页

[54]发明名称 利用文本的语义表示进行信息检索

[57]摘要

本发明涉及利用文本的语义表达进行信息检索。在一种优选实施例中,记号化器从输入字符串生成表征该输入字符串中所表达的语义关系的信息检索记号。记号化器首先从输入字符串建立表征输入字符串中的选定词之间的语义关系的原逻辑形式。接着记号化器确定和输入字符串中选定词之一具有"isa(是)"关系的超属词。然后记号化器从原逻辑形式构造一个或多个替代逻辑形式。通过为输入字符串中的一个或多个选定词中的每个选定词用为该选定词确定的某超属词代替原逻辑形式中的该选词,记号化器构造各替代逻辑形式。最后,记号化器生成代表原逻辑形式和替代逻辑形式的记号。最好使用记号化器生成记号,以便既用于构造代表目标文档的索引又用于对照索引处理查询。



ISSN1008-4274

权 利 要 求 书

1. 计算机系统中一种用于从输入字符串生成信息检索记号的方法，该方法包括步骤:

从输入字符串建立表征该输入字符串中选定的词之间的语义关系的原逻辑形式;

确定该输入字符串中各选定词的超属词;

从该原逻辑形式构造一个或多个替代的逻辑形式，通过为该输入字符串中的一个或多个选定词中的每个词用对该选定词确定的超属词代替原逻辑形式中的该选定词，构造每个替代的逻辑形式; 以及

生成代表原逻辑形式以及替代逻辑形式的记号，所生成的记号可由信息检索引擎区分。

2. 权利要求 1 的方法，其中构造步骤包括对输入字符串进行语法分析以判明其语法及语义结构的步骤.

3. 权利要求 1 的方法，其中确定步骤包括步骤:

对输入字符串中的每个选定词:

从语言知识库中检索该选定词的一个或多个超属词，每个超属词具有一个表征该超属词对该选定词在含义上的相似性的相似性值; 以及

确定其相似性值超过某预先建立的阈值的所有超属词。

4. 权利要求 1 的方法，还包括步骤:

在构造步骤之前，从某搜索查询选择输入字符串; 以及

把生成的记号提交给查询引擎以和一份或多份目标文档的表达进行比较。

5. 权利要求 1 的方法，还包括步骤:

在构造步骤之前，从要编排索引的文本体中选择输入字符串; 以及

把生成的记号提交给索引子系统以存储在代表该文本体的索引中。

6. 权利要求 5 的方法，还包括确定替代逻辑形式中出现的每个词的文档频率倒数的步骤，并且其中提交步骤不向索引子系统提交这样的表示替代逻辑形式的记号，即这些逻辑形式所包含的词的文档频率倒数

小于预先确定的最小文档频率倒数。

7. 权利要求 5 的方法，还包括步骤：

在提交步骤之后，确定替代逻辑形式中出现的每个词的文档频率倒数；以及

从索引中去掉这样的表示替代逻辑形式的记号，即这些逻辑形式所包含的词的文档频率倒数小于预先确定的最小文档倒数。

8. 权利要求 1 的方法，其中确定步骤确定相对于选定词具有相干亚属词集的选定词的超属词。

9. 一种计算机可读介质，其内容使计算机系统通过执行下述步骤从输入字符串中生成信息检索记号：

从输入字符串建立表征该输入字符串中选定的词之间的语义关系的原逻辑形式，

确定该输入字符串中各选定词的超属词；

从该原逻辑形式构造一个或多个替代的逻辑形式，通过为该输入字符串中的一个或多个选定词中的每个词用对该选定词确定的超属词代替原逻辑形式中的该选定词，构造每个替代的逻辑形式；

生成代表原逻辑形式以及替代逻辑形式的记号，所生成的记号可由信息检索引擎区分。

10. 权利要求 9 的计算机可读介质，其中构造步骤包括对输入字符串进行语法分析以判明其语法及语义结构的步骤。

11. 权利要求 9 的计算机可读介质，其中确定步骤包括步骤：

对输入字符串中的每个选定词：

从语言知识库中检索该选定词的一个或多个超属词，每个超属词具有一个表征该超属词对该选定词在含义上的相似性的相似性值；以及

确定其相似性值超过某预先建立的阈值的所有超属词。

12. 权利要求 9 的计算机可读介质，其中该计算机可读介质的内容还使计算机系统执行步骤：

在构造步骤之前，从某搜索查询选择输入字符串；以及

把生成的记号提交给查询引擎以和一份或多份目标文档的表达进行

比较。

13. 权利要求 9 的计算机可读介质，其中该计算机可读介质的内容还使计算机系统执行步骤：

在构造步骤之前，从要编排索引的文本体中选择输入字符串；以及

把生成的记号提交给索引子系统以存储在代表该文本体的索引中。

14. 一种计算机存储器，含有表征一份或几份目标文档的内容的文档索引数据结构，该文档索引数据结构把词映射到目标文档中的位置，该文档索引数据结构为各目标文档中出现的多个词段中的每个词段，把从该词段生成的逻辑形式中所包含的各个词映射到与该词段相对应的位置上，并且把从该词段生成的逻辑形式中所包含的各词的超属词映射到与该词段相对应的位置上，从而可把该文档索引数据结构用于响应接收到查询确定出目标文档中语义上类似于查询段的词段位置。

15. 权利要求 14 的计算机存储器，其中文档索引数据结构把至少一个未在任一目标文档中出现的词映射到目标文档的某位置上。

16. 一种用于响应查询的计算机系统，查询包含着与一份或多份目标文档对照的词段，每份目标文档包含一个或多个词段，每个目标文档段具有目标文档中的一个位置，该计算机系统包括：

目标文档接收器，用于接收目标文档；

查询接收器，用于接收对各目标文档的查询；

记号化器，用于从目标文档接收器接收到的目标文档的词段以及从查询接收器接收的查询生成记号，该记号化器包括用于从每个词段合成出一个表征该词段的语义结构的逻辑形式的逻辑形式合成器，该记号化器生成代表从词段中合成出的逻辑形式的记号；

索引存储器，用于存储把每个从某目标文档段生成的记号映射到生成该记号的目标文档段在目标文档中的位置上的关系；以及

查询处理子系统，用于为每次查询在索引存储器中确定和从该查询生成的记号匹配的某记号，并用于返回从该确定的记号映射到的位置的指示。

17. 权利要求 16 的计算机系统，其中逻辑形式合成器合成的逻辑形式包含若干词，并且记号化器还包括：

超属词扩展子系统，用于从逻辑形式合成器生成的逻辑形式创造一个或多个用超属词替代该逻辑形式中的一个或多个词的辅助的逻辑形式，记号化器还生成代表由超属词扩展子系统创造的辅助逻辑形式的记号。

说　　明　　书

利用文本的语义表示进行信息检索

本发明涉及信息检索领域，并且更具体地涉及信息检索记号化领域。

信息检索指的是确定目标文档中出现查询或查询文档中的词的过程。信息检索可以被有益地应用于几种情况中，包括：处理用户的明确搜索查询，确定和某特定文档相关的文档，判断两份文档的类似性，提取某文档的特征以及概述某文档。

信息检索典型地包括两阶段过程：（1）在编索引阶段，最初通过（a）把文档中的每个词转化成信息检索引擎可理解、可区分的一串字符，称之为“记号”（即　文档的记号化）以及（b）建立各记号到该记号在该文档中出现位置的索引，对文档编索引。（2）在查询阶段中，相似地对查询（或查询文档）进行记号化，并和索引进行比较以确定文档中出现记号化后的查询中的记号的位置。

图 1 是描述信息检索过程的概述数据流图。在编索引阶段，把目标文档 111 提供给记号化器 112。目标文档是由一些字符串，例如一些句子，组成的，每个字符串出现在目标文档的某特定位置上。将目标文档中的各字符串以及词的位置传送到记号化器 120，记号化器 120 把各字符串中的词转换成一系列可由信息检索引擎 130 理解及区分的记号。信息检索引擎 130 的索引建立部分 131 把这些记号以及它们的位置添加到索引 140 中。该索引把每个唯一的记号映射到该目标文档中出现该记号的位置。若需要，可以重复该过程，以便把一些不同的目标文档添加到该索引中。若索引 140 表示一些目标文档中的文本，则位置信息最好包含各位置对应的文档的标记。

在查询阶段，把文本查询 112 提供给记号化器 120。查询可能是单个字符串或一个句子，或者可能是由一些字符串组成的完整文档。记号化器 120 按它把目标文档中的词转换成记号的相同方式把查询 112 的文

本中的词转换成记号。记号化器 120 把这些记号传送到信息检索引擎 130 的索引检索部分 132。信息检索引擎的索引检索部分在索引 140 中搜索这些记号在目标文档中的出现。对于每个记号，信息检索引擎的索引检索部分确定目标文档中出现该记号的各个位置。作为查询结果 113 返回位置表。

常规记号化器典型地包括输入文本的外表变换，例如把每个大写字符变成小写、确定输入文本中的每个词并且去掉词的后缀。例如，常规记号化器可能把输入的文本字符串

The father is holding the baby.

（该 父亲 正抱着 该 婴儿.）

转换成下述记号:

the （该）
father （父亲）
is （是）
hold （抱）
the （该）
baby （婴儿）

这种记号化方法趋向于使依据它的搜索过分地包含出现这样的词，即其含意是和查询文本中的预定含意不同的。例如，该示例输入文本字符串使用“to support or grasp（支持或抓住）”含意下的动词“hold”。但是，记号“hold”可能会和其含意是“the cargo area of a ship（船的装货区”）的词“hold”匹配。这种记号化方法还趋向于过分包含这样的情况，即其中词之间的关系和查询文本中各词之间的关系不同。例如，在上述示例输入文本字符串中，“father”是词“hold”的主语而“baby”是宾语，该示例的字符串可能和句子“The father and the baby held the toy”匹配，在该句中，“baby”是主语而不是宾语。该方法还会过少地包括出现这样的情况，即采用不同的但在语义上相关的词来代替查询文本中的某个词。例如，上述的输入文本字符串可能不和文本字符串“The parent is holding the baby”匹配。出于常规记号化方法的这些缺点，一种编有记号化文本中隐含的语义关系的记号化器应该是非常实用的。

本发明目的是利用一种改进的记号化器进行信息检索，该改进的记号化器分析输入文本以确定逻辑形式，接着利用超属词扩展逻辑形式。当和常规信息检索索引结构以及查询一起使用时，本发明减少标识出现不同的含意以及标识出现词之间带有不同的关系的次数，并且增加标识出现使用不同的但在语义上相关的用语的次数。

通过对已编索引的文本和查询文本进行语法分析以对该输入文本进行词法、语法和语义分析，本发明克服了和常规记号化过程相关的问题。该分析过程产生一个或多个逻辑形式，它们标识查询文本中起主要作用的词以及它们预定的含意，并且还进而确定这些词之间的关系。该语法分析程序最好产生和输入文本的深主语、动词和深宾语相关的逻辑形式。例如，对于输入文本“The father is holding the baby”，语法分析程序可能生成下述逻辑形式:

深主语	动词	深宾语
father	hold	baby

语法分析程序还将该输入文本中采用的特定含意归入这些词。

利用数字词典或辞典（也称为语言知识库）为某词的某特定含义确定和该词的该含义为通用术语的其它词的含义（“超属词”），本发明把语法分析程序生成的逻辑形式中的词改变成它们的超属词以创造附加的逻辑形式，这些附加的逻辑形式所具有的总含义和原始逻辑形式的含义相接近。例如，根据词库中的指示，“parent”的一种含意是“father”的所属含意的超属词，“touch”的一种含意是“hold”的所属含意的超属词，“child”的一种含意以及“person”的一种含意是“baby”的所属含意的超属词，本发明可建立如下的附加逻辑形式:

深主语	动词	深宾语
parent	hold	baby
father	touch	baby
parent	touch	baby
father	hold	child
parent	hold	child
father	touch	child

parent	touch	child
father	hold	person
parent	hold	person
father	touch	person
parent	touch	person

然后，本发明把所有生成的逻辑形式变换成可由信息检索系统理解的记号，该系统把记号化后的查询和索引进行比较，并且提供给该信息检索系统。

图 1 是信息检索过程的概述数据流图.

图 2 是最好在其上运行本工具的通用计算机系统的高级框图.

图 3 是一个概述流程图，表示最好由本工具执行的各步骤以便构造和访问语义上代表目标文档的索引.

图 4 是一个流程图，表示由本工具使用的用以生成输入句子的各记号的记号化例程.

图 5 是一个逻辑形式图，表示示例的逻辑形式.

图 6 是一个输入文本图，表示输入文本片断，本工具为这些片断构造图 5 中示出的逻辑形式.

图 7A 是一个语言知识库图，表示由语言知识库确定的示例性超属词关系.

图 7B 是一个语言知识库图，表示为原逻辑形式的深主语 man（含意 2）选择超属词.

图 8 是一个语言知识库图，表示为原逻辑形式的动词 kiss（含意 1）选择超属词.

图 9 和 10 是语言知识库图，表示为原逻辑形式的深宾语 pig（含意 2）选择超属词.

图 11 表示扩展逻辑形式的逻辑形式.

图 12 表示通过置换扩展的原逻辑形式建立派生的逻辑形式.

图 13 是一个索引图，表示索引内容的例子.

图 14 是一个逻辑形式图，表示本工具为查询“man kissing horse”优选构造的逻辑形式.

图 15 表示利用超属词扩充原逻辑形式.

图 16 是一个语言知识库图，表示选择查询逻辑形式的深宾词 horse（含意 1）的超属词。

图 17 是部分逻辑形式图，表示和一个只包含深主语和动词的部分查询对应的部分逻辑形式。

图 18 是部分逻辑形式图，表示和一个只包含动词和深宾语的部分查询对应的部分逻辑形式。

本发明的目的是利用文本的语义表达进行信息检索。当和常规信息检索索引结构以及查询一起使用时，本发明减少标识出现不同的含意以及标识出现词之间存在不同的关系的次数，并且增加标识出现使用不同的但在语义上相关的用语的次数。

在一种优选实施例中，用一种改进的信息检索记号化工具（以下称“本工具”）代替图 1 所示的常规记号化器，该工具分析输入文本以确定逻辑形式，接着利用超属词扩展逻辑形式。通过对已编索引的文本和查询文本进行语法分析以对该输入文本进行词法、语法和语义分析，本发明克服了和常规记号化过程相关的问题。该分析过程产生一个或多个逻辑形式，它们标识查询文本中起主要作用的词以及它们的预定含意，并且还进而确定这些词之间的关系。该语法分析程序最好产生和输入文本的深主语、动词和深宾语相关的逻辑形式。例如，对于输入文本“The father is holding the baby”，该语法分析程序可产生表示深主语是“father”、动词是“hold”及深宾语是“baby”的逻辑形式。由于把输入文本转换成逻辑形式通过去掉修饰语并忽略时态和语态的差异将输入文本“蒸馏”成基本含义，把输入文本片断转换成逻辑形式趋于统一自然语言中表达相同思想可能采用的许多不同方式。该语法分析程序还确定这些词在该输入文本中所使用的特定含义。

利用数字词典或辞典（也称为语言知识库）为某词的某特定含义确定和该词的该含义为通用术语的其它词的含义（“超属词”），本发明把语法分析程序生成的逻辑形式中的词改变成它们的超属词以创造附加的逻辑形式，这些附加的逻辑形式所具有的总含义和原始逻辑形式的含义相接近。然后，本发明把所有生成的逻辑形式变换成可由信息检索系统理

解的记号，该系统把记号化后的查询和索引进行比较，并且提供给该信息检索系统。

图 2 是最好在其上运行本工具的通用计算机系统的高级框图。计算机系统 200 包括中央处理器（CPU）210、输入/输出部件 220 及计算机存储器（存储器）230。输入/输出部件中有存储部件 221，例如硬盘机。输入/输出部件还包括计算机可读的介质驱动器 222，它可用于安装软件产品，其中包括计算机可读介质如 CD－ROM 上提供的本工具。输入/输出部件还包括因特网连接 223，其使计算机系统 200 通过因特网和其它计算机系统通信。最好包括本工具 240 的计算机程序驻留在存储器 230 中并在 CPU 210 上执行。本工具 240 包括一个基于规则的语法分析程序，用于分析要记号化的输入文本片断以生成逻辑形式。本工具 240 还包括一个由该语法分析程序使用的语言知识库 242，以把含义号赋予逻辑形式中的词。本工具还利用语言知识库确定所生成的逻辑形式中的各词的超属词。存储器 230 最好还包括索引 250，其用于将根据目标文档生成的记号映射到目标文档中的位置。存储器 230 还包括一个信息检索引擎（“IR 引擎”）260，用于把从目标文档生成的记号存储到索引 250 中，并且用于确定索引中和从查询生成的记号相匹配的记号。尽管本工具最好在按上述配置的计算机系统中实现，熟练技术人员可意识到它可实现在具有不同配置的计算机系统上。

图 3 是一个概述流程图，表示为了构造和访问语义上代表目标文档的索引最好由本工具执行的步骤。简言之，本工具首先通过把目标文档的每个句子或句子片断变换成一些记号在语义上对目标文档编索引，这些记号表示描述句子中重要的词之间的关系的扩展逻辑形式，并包括着具有类似含义的超属词。本工具把这些“语义记号”以及目标文档中出现该句子的位置存储到索引中。当对所有目标文档编排索引后，本工具能对照该索引处理信息检索查询。对于接收到的每条这种查询，本工具以对来自目标文档的句子进行记号化的相同方式对查询文本记号化－即通过把句子变换成共同表示查询文本之扩展逻辑形式的各语义记号。然后，本工具把这些语义记号和索引中存储的语义记号进行比较，以确定目标文档中存储的这些语义记号的位置，并且按照与该查询的关联顺序

对包含这些语义记号的目标文档分类。本工具最好可更新索引，以便随时包含新目标文档的语义记号。

参照图 3，在步骤 301－304，本工具循环处理目标文档中各个句子。在步骤 302，本工具调用例程以记号化图 4 所示的句子。

图 4 是一个流程图，表示本工具使用的生成输入句子或其它输入文本片断的记号的记号化例程。在步骤 401，本工具从输入文本片断构造原逻辑形式。如上面所讨论。逻辑形式表示句子或句子片断的基本含义。通过应用语法分析程序 241（图 2）使输入文本片断得到语法及语义分析处理产生逻辑形式。对于构造表示输入文本字符串的逻辑形式的详细讨论，请参见美国专利申请 08/674，610 号，这里引用作为参考。

本工具使用的逻辑形式最好析出句子的主要动词、该动词的实际主语的名词（“深主语”）以及该动词的实际宾语的名词（“深宾语”）。图 5 是一个逻辑形式图，表示示例的原逻辑形式。该逻辑形式具有三个元素“深主语元素 510、动词元素 520 以及深宾语元素 530。可以看出，该逻辑形式的深主语是词“man”的含义 2。含义号为具有多于一个含义的词指示语法分析程序赋予词的特定含义，该含义是由语法分析程序所使用的语言知识库定义的。例如，词“man”可具有意思为人的第一含义和具有成年男性的第二含义。逻辑形式的动词是词“kiss”的第一含义。最后，深宾语是词“pig”的第二含义。该逻辑形式的简化版本是一个有序三元组 550，其第一元素是深主语，第二元素是动词，其第三元素是深宾语：

（man, kiss, pig）

图 5 中所示的逻辑形式表征一些不同的句子和句子片断。例如，图 6 是一个表示输入文本片断的输入文本图，本工具会为其构造图 5 中所示的逻辑形式。图 6 表示输入文本句子片断“man kissing a pig”。可以看出该短语出现在文档 5 的词号 150 处，占据着词位置 150、151、152 和 153。当本工具对该输入文本蒸片断进行记号化时，它生成图 5 中示出的逻辑形式。本工具也会为下述输入文本片断生成图 5 中所示的逻辑形式：

The pig was kissed by an unusual man.

The man will kiss the largest pig.

Many pigs have been kissed by that man.

如前面所讨论，由于把输入文本转换成逻辑形式通过去掉修饰语并忽略时态和语态的差异将输入文本蒸馏成基本含义，把输入文本片断转换成逻辑形式趋于统一自然语言中表达相同思想可能采用的许多不同方式。

回到图 4，在本工具从输入文本构造出原逻辑形式后，例如图 5 中所示的逻辑形式后，本工具进入步骤 420 以利用超属词扩展该原逻辑形式。在步骤 402 后，记号化例程返回。

如上面所述，超属词是一个属术语，它和某特定的词具有“is a ”（是）的关系。例如，词“vehicle”是词“automobile”的超属词。本工具最好利用一个语言知识库确定原逻辑形式下的词的超属词。这种语言知识库典型地包含规定某词的超属词的语义链接。

图 7A 是一个语言知识库图，表示由语言知识库确定的示例超属词关系。请注意，类似于后面的语言知识库，图 7A 已被简化以便利本说明，并且略掉通常可在语言知识库中发现的不和本说明直接相关的信息。图 7A 中的每个向上的箭头把某个词和它的超属词连接起来。例如，有一个箭头把词 man（含义 2）711 连接到词 person（含义 1）714，表示 person（含义 1）是 man（含义 2）的超属词。相反，man（含义 2）被说成是 person（含义 1）的“亚属词”。

在为了扩展原逻辑形式而确定超属词时，本工具根据超属词的亚属词的相关为原逻辑形式的每个词选择一个或多个超属词。通过以这种方式选择超属词，本工具在超出输入文本片断含义的范围外（但在控制量内）使逻辑形式的含义广义化。对于某原逻辑形式中的某特定词，本工具首先选择该原逻辑形式的该词的直接超属词。例如，参照图 7A，从原逻辑形式中的 man（含义 2）711 开始，本工具选择它的超属词 person（含义 1）714。下一步，本工具根据 person（含义 1）714 是否具有相对于起始词 man（含义 2）711 的相关亚属词集，判定是否还要选择 person（含义 1）714 的超属词 animal（含义 3）715。若与起始词 man（含义 2）711 不同的词 person 的所有含义的大量亚属词至少具有对起始词 man（含义 2）711 的相似性的临阈级，则 person（含义 1）714 具有相对于 man（含义 2）711 的相干亚属词集。

为了确定超属词的不同含义的亚属词之间的相似度，本工具最好咨询语言知识库以得到表示词的这些词句之间的相似程度的相似性权重。图 7B 是一个语言知识库图，表示 man（含义 2）和 person（含义 1）的及 person（含义 5）的其它亚属词之间的相似性权重。该图表示：man（含义 2）和 woman（含义 1）之间的相似性加权是“.0075”；在 man（含义 2）和 child（含义 1）之间的相似性权重是“.0029”；在 man（含义 2）和 villain（含义 1）之间的相似性权重是“.0003”；以及在 man（含义 2）和 lead（含义 7）之间的相似性权重是“.0002”。这些相似性加权最好是由语言知识库根据该语言知识库保持的词意对之间的语义关系网络计算的。关于利用语言知识库计算词义对之间的相似性加权的详细讨论，请参见标题为“确定词之间的相似性”的美国专利申请　　号（专利律师卷号 661005.524），这里引用作为参考。

为了根据这些相似性加权判定亚属词集是否相干，本工具确定相似性加权的阈值量是否超过相似性加权阈。虽然优选阈百分比是 90%，最好为了优化本工具的性能调整阈百分比。还可把相似性加权阈值配置成优化本工具的性能。相似性加权阈值最好和语言知识库提供的相似性加权的总分布相配合。这里，示出采用“.0015”的阈值。从而本工具判定起始词的和超属词的所有含义的其它亚属词之间的至少 90% 的相似性加权是否等于或高于“.0015”的相似性加权阈。可以从图 7B 看出，相对于 man（含义 1）的 person 的亚属词不满足该条件：尽管 man（含义 1）和 women（含义 1）之间以及 man（含义 1）和 child（含义 1）之间的相似性加权大于“.0015”，man（含义 1）和 villain（含义 1）之间以及 man（含义 1）和 lead（含义 7）之间的相似性加权小于“.0015”。从而本工具不再选择超属词 animal（含义 3）715，也不选择 animal（含义 3）的任何超属词。因此，只选择超属词 person（含义 1）714 用于扩展原逻辑形式。

为了扩展原逻辑形式，本工具还选择原逻辑形式的动词和深宾语的超属词。图 8 是一个语言知识库图，表示选择原逻辑形式的动词 kiss（含义 1）的超属词。从图中可看出 touch（含义 2）是 kiss（含义 1）的超属词。该图还示出 kiss（含义 1）和 touch 的所有含义的其它亚属词之间

的相似性加权. 本工具首先选择原逻辑形式的动词 kiss（含义 1）的直接超属词 touch（含义 2）. 为了判定是否选择 touch（含义 2）的超属词 interact（含义 9），本工具判定 kiss（含义 1）和 touch 的所有含义的其它亚属词之间的相似性加权中有多少至少和相似性加权阈值一样大。由于这四个相似性加权中只有两个至少和“.0015”的相似性加权阈值一样大，所以本工具不选择 touch（含义 2）的超属词 interat（含义 9）.

图 9 和图 10 是语言知识库图，表示选择原逻辑形式的深宾语的超属词和 pig（含义 2）. 从图 9 中可以看出本工具选择 pig（含义 2）的超属词 swine（含义 1）和选择 swine（含义 1）的超属词 animal（含义 3）来扩展原逻辑形式，因为 swine 的唯一含义的 90% 以上（事实上，100%）的超属词具有等于或高于“.0015”的相似性加权阈值. 从图 10 中可以看出，本工具不继续选择 animal（含义 3）的超属词 organism（含义 1），因为 animal 的含义的超属词中具有等于或高于“.0015”相似性加权阈值的超属词少于 90%（实际上 25%）.

图 11 是一个逻辑形式图，表示扩展逻辑形式。从图 11 中可以看出，扩展逻辑形式的深主语元素 1110 包括除词 man（含义 2）1111 之外的超属词 person（含义 1）. 可看出动词元素 1120 包括超属词 touch（含义 2）1112 和词 kiss（含义 1）1121。还可以看出，扩展逻辑形式的深宾语包括除词 pig（含义 2）1131 之外的超属词 swine（含义 1）和 animal（含义 3）1132.

通过在扩展逻辑形式的各个元素中用超属词置换原始词，本工具可创造一个数量比较大的派生逻辑形式，这些逻辑形式在意义上和原逻辑形式比较接近。图 12 表示通过置换扩展的原逻辑形式建立的派生逻辑形式。从图 12 中可看出，此置换创造十一个派生逻辑形式，每个逻辑形式在比较准确的方式下表征输入文本的含义。例如，图 12 示出的派生逻辑形式。

（person, touch, pig）

在含义上非常接近句子片断

man kissing a pig

图 11 中所示的扩展逻辑形式表示原逻辑形式加这十一个派生逻辑形式，

它们被更紧凑地表示成扩展逻辑形式 1200:

((man OR person), (kiss OR touch), (pig OR swine OR animal))

本工具以允许记号可由常规信息检索引擎处理的方式，从该扩展逻辑形式生成逻辑记号。首先，本工具把某保留字符附加到扩展逻辑形式中的各个词上，以确定输入文本片断中出现的词是否是深主语、动词或深宾语。这可确保，当词“man”作为深主语出现在查询输入文本的扩展逻辑形式中时，它不会和存储在索引中的作为动词出现在某扩展逻辑形式的一部分的词“man”匹配。一将保留字符映射为逻辑格式元素的示例 如下:

逻辑形式元素	标识字符
深主语	-
动词	^
深宾语	#

利用保留字符的这种示例映射，为逻辑形式“(man, kiss, pig)”生成的记号应包括“man_ ”，“kiss^”以及“pig # ”。

常规信息检索引擎生成的索引通常把每个记号映射到目标文档中出现该记号的各特定位置。常规信息检索引擎可能利用文档号和词号表示这种目标文档位置，文档号标识包含着该记号的目标文档，词号标识该目标文档中出现该记号的位置。这种目标文档位置允许常规信息检索引擎确定在目标文档中一起出现的多个词，以响应利用“PHRASE (短语)”运算符的查询，该运算符要求其联接的词在目标文档中是相邻的。例如，查询“red PHRASE bicycle”将匹配出现在文档 5 词 611 处的“red”以及在文档 5 词 612 处的“bicycle”，但不会匹配出现在文档 7 词 762 处的“red”以及在文档 7 词 202 处的“bicycle”。把目标文档位置存储在索引中还允许常规信息检索引擎响应查询确定目标文档中出现被查询记号的各个点。

对于来自目标文档输入文本片断的扩展逻辑形式，本工具最好类似地向每个记号分配人工目标文档位置，即使扩展逻辑形式的这些记号实际上并不在目标文档中的这些位置上出现。分配这些目标文档位置既 (A) 允许常规搜索引擎利用 PHRASE 运算符确定和单个原逻辑形式或派生逻

辑形式对应的语义记号的组合，又（B）允许本工具把分配的位置和目标文档中的输入文本片断的实际位置关联起来。从而本工具按如下向语义记号分配位置.

逻辑形式元素	位置
深主语	（输入文本片断中第 1 个词的位置）
动词	（输入文本片断中第 1 个词的位置）+ 1
深宾语	（输入文本片断中第 1 个词的位置）+ 2

从而本工具按如下对从文档 5、字 150 处开始的句子得到的“(man, kiss, pig)” 的扩展逻辑形式的记号分配目标文档位置：“man_” 和“person_” ——文档 5，词 150；“kiss^” 和“touch^” ——文档 5，词 151；以及“pig#”、“swine#” 和“animal#” ——文档 5，词 152。

回到图 3，在步骤 303，本工具把记号化例程建立的记号以及它们的出现位置存储到索引中。图 13 表示索引的示例内容。索引将每个记号映射到文档的标识上以及该记号在该文档中的出现位置。请注意，尽管索引是作为表示出的，以便更清楚地表示索引中的映射，实际上最好把索引存储到一些其它的更有效支持索引中的记号的位置的格式中的一种格式中，例如树状格式。另外，最好利用诸如前缀压缩技术压缩索引中的内容，以将索引的长度降到最低限度。

可以看出，根据步骤 303，本工具为扩展逻辑形式下的各个词的索引 1300 中存储了映射。在索引中存储了从深主语词“man” 和“person” 到文档号 5、词号 150 处的目标文档位置的映射。词号 150 是在该处开始图 6 中所示的输入文本片断的词位置。可以看出，本已把保留字符“_” 附加在和深主语词对应的记号上。通过附加该保留字符，当以后搜索该索引时，本工具能检索这些词作为逻辑形式的深主语出现的情况，而不检索这些词作为逻辑形式的动词或深宾语的出现。类似地，该索引包括动词“kiss” 和“touch” 的记号。这些动词词的条目把它们映射到文档 5、词号 151 的目标文档位置上，即深主语词的目标文档位置的后一个词。还可以看出，已为这些动词词的记号附加了保留字符“^”，从而这些词的出现以后不会作为深主语或深宾语元素出现。类似地，该索引包含深宾语词“animal”、“pig” 和“swine” 的记号，把它们映射到文档号 5、

词号 152 的目标文档位置上，即该短语开始的目标文档位置的两个词后。对深宾语词的记号附加保留字符“#”以把它们标识为索引中的深宾语。利用以这种状态示出的索引，通过搜索图 12 示出的任一派生原逻辑形式的索引，可以找到图 6 中所示的输入文本片断。

在一种优选实施例中，本工具在同一索引中存储目标文档中字面上出现的词到其目标文档中的实际位置的映射以及该目标文档的语义表达，最好用一个常数递增语义表达的各个语义记号的词号值，其中该常数大于任一文档中的词的数量，以便在访问该索引时把语义表达的语义记号和文字记号区分开来。为了简化图 13，未示出添加该常数。

在该例子，本工具将扩展逻辑形式中的每个词的记号添加到索引中，以形成目标文档的语义表达。然而，在一种优选实施例中，本工具对那些可能在区分各目标文档中的文档是有效的逻辑形式记号，限制添加到索引中的扩展逻辑形式记号集。为了如此限制添加剂索引的扩展逻辑形式记号集，本工具最好确定各记号文档频率倒数，其公式由后面的式（1）表示。在该实施例，本工具只把其文档频率倒数超过最小阈值的记号添加到索引中。

回到图 3，在目标文档的当前句子之前把记号存储到索引中后，在步骤 304，本工具循环回到步骤 301 以处理目标文档中的下个句子。当处理完目标文档中的所有句子时，本工具进入步骤 305。在步骤 305，本工具接收查询文本。在步骤 306－308，本工具处理接收到的查询。在步骤 306，本工具调用记号化例程以对查询文本记号化。图 14 是一个逻辑形式图，表示根据步骤 401（图 4）最好由本工具为查询“man kissing horse”构造的逻辑形式。可以该逻辑形式图中看出，深主语是 man（含义 2），动词是 kiss（含义 1），深宾语是 horse（含义 1）。该原逻辑形式更简明地表达成原逻辑形式 1450。

（man, kiss, horse）

图 15 表示根据步骤 402（图 4）利用超属词扩展原逻辑形式，从图 15 可看出，类似于取自目标文档的示例输入文本，用超属词 person（含义 1）扩展深主语 man（含义 2），用超属词 touch（含义 2）扩展动词 kiss（含义 1），还可以看出，用超属词 animal（含义 3）扩展深宾语 horse

（含义 1）.

图 16 是一个语言知识库图，表示选择查询逻辑形式的深宾语 horse（含义 1）的超属词. 从图 16 中可以看出，由于 animal（含义 3）的亚属词中少于 90% 的亚属词具有的相似性加权等于或高于“.0015”的相似性加权阈值，所以本工具不选择 animal（含义 3）的超属词 organism（含义 1）. 从而，本工具只利用超属词 animal（含义 3）扩展逻辑形式.

回到图 3，在步骤 307，本工具使用扩展逻辑形式 1550（图 15）检索目标文档中出现匹配记号的索引位置，该扩展逻辑形式 1550 是利用原逻辑形式的词含义的超属词构造的。本工具最好通过发出下述与索引对比的查询:

（man_OR person_） PHRASE (kiss ∧OR touch ∧) PHRASE (horse # OR animal #)进行检索。PHRASE 运算符匹配出现这样的情况，即，该运算符后的操作数的词位置 1 比其前面的操作数的词位置大。从而，该查询匹配在动词 kiss^或 touch^之前的深主语 man_或 person_，其中动词 kiss^或 touch^在深宾语 horse #或 animal # 之前。从图 13 的索引可看出，在文档号 5、词号 150 处满足该查询。

若该查询不满足该索引，则本工具将继续提出两个不同部分查询下的查询。第一个部分形式只包括深主语和动词，不包括宾语:

(man_ OR person_) PHRASE (kiss ∧ OR touch ∧)

图 17 是一个部分逻辑形式图，表示和该第一查询对应的部分逻辑形式。查询的第二部分形式包括动词和深宾语，但不包括深主语:

（kiss ∧ OR touch ∧）PHRASE (horse # OR animal #)

图 18 是一个部分逻辑形式图，表示和该第二部分查询对应的部分逻辑形式。这些部分查询会和索引中具有不同深主语或深宾语的逻辑形式匹配，并且会和不具有深主语或深宾语的部分逻辑形式匹配. 这些部分查询考虑查询输入文本片断和目标文档输入文本片断之间的差异，其中包括代词的使用以及暗含的深主语以及深宾语.

回到图 3，在确定索引中记号的匹配后，本工具进入步骤 308 以对目标文档分类，其中按它们与查询的关联性的顺序出现和原逻辑形式或派生逻辑形式对应的各匹配记号的特定组合的匹配. 在本发明的不同实

施例中，本工具采用一些周知方法中的一种或几种通过关联性对各文档分类，这些方法包括 Jaccard 加权和二进制项独立加权。本工具最好采用文档频率倒数和项频率等待的组合对匹配的目标文档分类。

在对目标文档中出现较少的记号组合给予较大的加权下，文档频率倒数加权表征记号组合区分文档的能力。例如，对于一组主题是 photography（摄影术）的一组目标文档，逻辑形式

（photographer, frame, subject）

会出现在该组文档中的每份文档中，从而对于区分各文档它不是一种很好的基准。由于上述逻辑形式在每份目标文档中出现，所以它具有较小的文档频率倒数。记号组合的文档频率倒数的公式如下：

$$\text{文档频率倒数（记号组合）} = \log\left(\frac{\text{目标资料总数}}{\text{含有记号组合的目标资料数}}\right) \qquad (1)$$

文档中记号组合的项频率加权量测该文档专用于该记号组合的程度，并假定其中多次出现某特定查询记号的文档要比在其中不太出现该查询记号的文档关联更大。文档中某记号组合的项频率加权公式如下：

项频率（记号组合，文档）= 该文档中出现该记号组合的次数 （2）

本工具利用各匹配文档的记分对文档分类。本工具首先利用下述公式对每份文档中的各匹配记号组合计算计分：

记分（记号组合，文档）

= 文档频率倒数（记号组合）× 项频率（记号组合，文档） （3）

接着本工具根据下式通过选择各匹配文档中任一匹配记号组合的最高记分，计算各匹配文档的记分：

$$\text{记分（文档）} = \max\left[\begin{array}{l}\forall \\ \text{资料中的} \\ \text{记号组合}\end{array}(\text{记分(记号组合，资料)})\right] \qquad (4)$$

一旦本工具计算出每份文档的记分，本工具可扩大这些记分以反映和那些指向语义匹配的项不同的查询项。在扩大每份文档的记分后，若需要，本工具通过按下式考虑文档的篇幅计算每份文档的归一化记分：

$$\text{归一化记分（文档）} = \frac{\text{记分（资料）}}{\text{篇幅（资料）}} \qquad (5)$$

篇幅（文档）项可以是某文档的篇幅的任何合理量测，例如该文档中的

字符、词、句子或句子片断的数量。可以替代地用一些其它归一化技术归一化文档记分，包括余弦测量归一化、项加权和归一化以及最大项加权归一化。

在计算出每份匹配文档的归一化记分后，本工具按文档的归一化记分的顺序对匹配文档分类。用户最好从分类表中选择一份匹配文档，以得到该文档中匹配记号组的位置，或者显示该文档的匹配部分。

回到图 3，在步骤 308 中对匹配的目标文档分类后，本工具最好进入步骤305以接收下个查询的文本以和索引对比。

上面讨论了通过关联性对包含匹配记号组的文档进行分类。本发明的其它优选实施例类似地通过关联性分别对包含匹配的文档集和文档段落分类。对于被组织成各包含一份或几份文档的文档集的目标文档，本工具最好通过关联性对出现匹配的文档集分类，以确定最相关的文档集供进一步查询。另外，本工具最好可配置成能把每份目标文档划成段落并且对其中出现匹配的文档段落的关联性分类。通过选择一数量的字节、词或句子或者使用目标文档中出现的结构、格式或语言线索，在目标文档中相邻标识这些文档段落。本工具最好还确定论及特定论题的不相邻的文档段落。

虽然参照各优选实施例显示并说明了本发明，熟练技术人员理解，在不背离本发明的范围下在形式和细节上可作出各种更改或修改。例如，记号化程序可以直接采纳或生成对应于一个完整的逻辑形式结构的记号以替代对应于某逻辑形式结构中的一个词的记号，并且把这样的记号存储到索引中。而且，可以应用各种周知技术以在具有语义匹配成分的查询中包括其它类型的搜索。并且，查询可包括若干语义匹配成分。此外，可利用标识词之间的语义关系代替超属词来扩展原逻辑形式。本工具还可以利用原逻辑形式的每个词的预先编译的替代词表扩展原逻辑形式，而不是如前面所说明的那样在运行时根据语言知识库生成超属性表。此外，为了提高匹配精度，记号化程序可以在词的记号中编码标识该词的含义号。在这种情况下，对超属词集的相干性的检查减少成不必为选定超属词的所有含义检查相似性。在本例中，只有词 person 的含义 1 的超属词需要带有对于词 man（含义 2）的起始含义的相似性阈值。由于索

引表中的可能匹配项歧义较少，我们可以限制可能产生的错误命中的项集．由于这个原因，只需要检查和逻辑形式中的词具有超属词关系的那些含义．

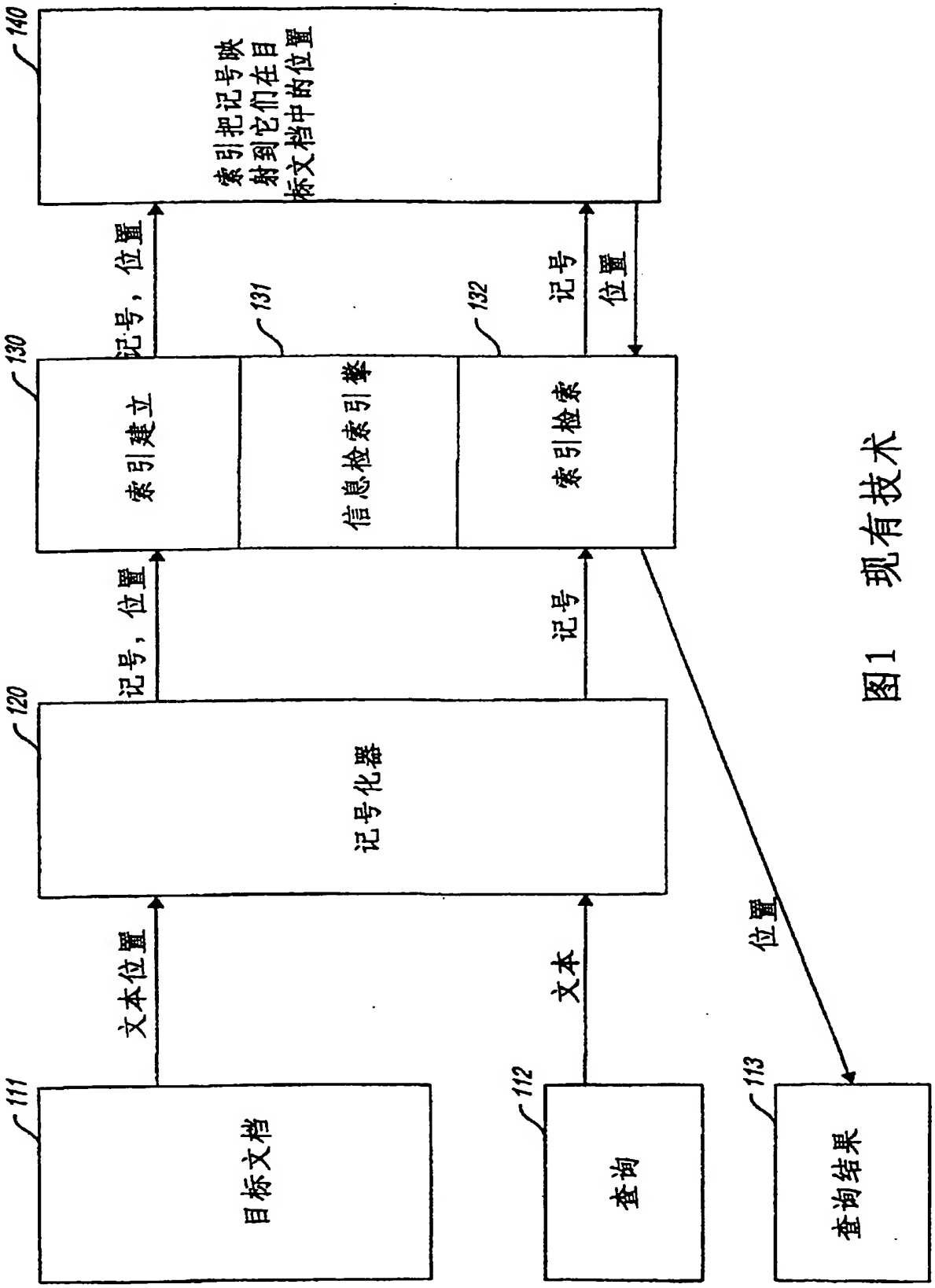

140
索引把记号映射到它们在目标文档中的位置
记号，位置
131
132
记号
位置
130
索引建立
信息检索引擎
索引检索
记号，位置
记号
120
记号化器
文本位置
文本
位置
111
目标文档
112
查询
113
查询结果

图1 现有技术

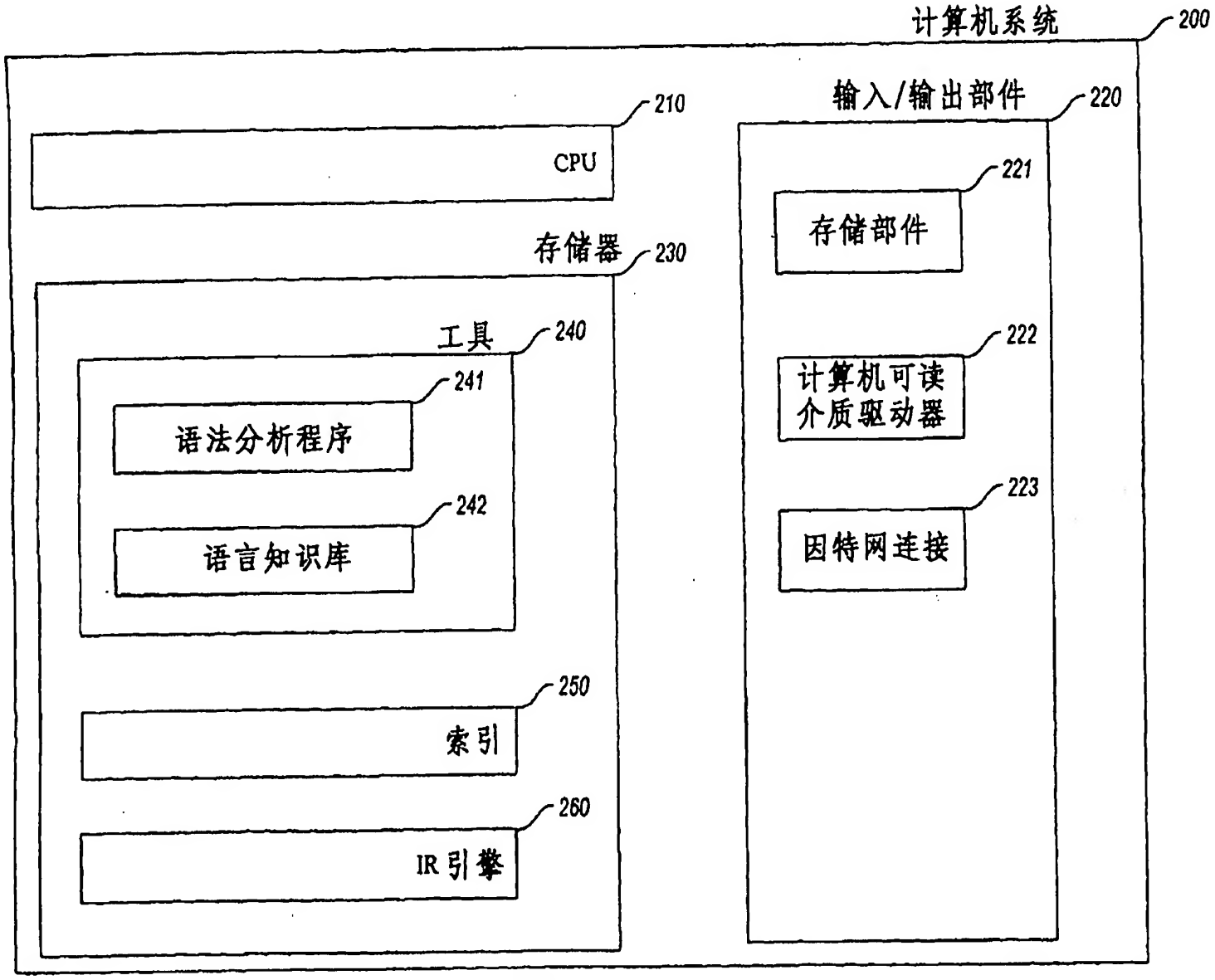

计算机系统
200
输入/输出部件
220
210
CPU
221
存储部件
存储器
230
222
计算机可读
介质驱动器
工具
240
241
语法分析程序
223
因特网连接
242
语言知识库
250
索引
260
IR 引擎

图 2

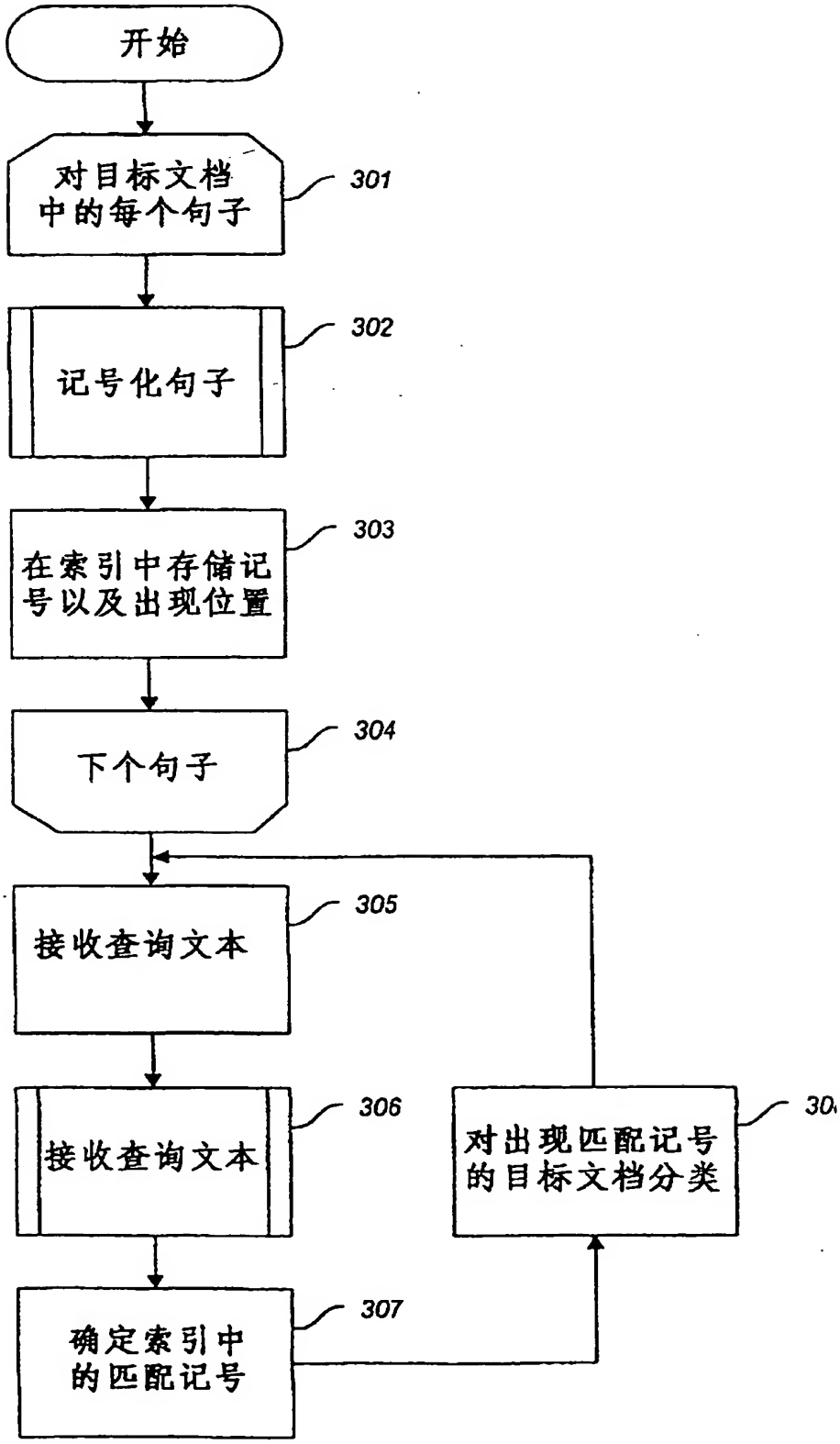

开始
对目标文档中的每个句子
301
记号化句子
302
在索引中存储记号以及出现位置
303
下个句子
304
接收查询文本
305
接收查询文本
306
对出现匹配记号的目标文档分类
确定索引中的匹配记号
307

图3

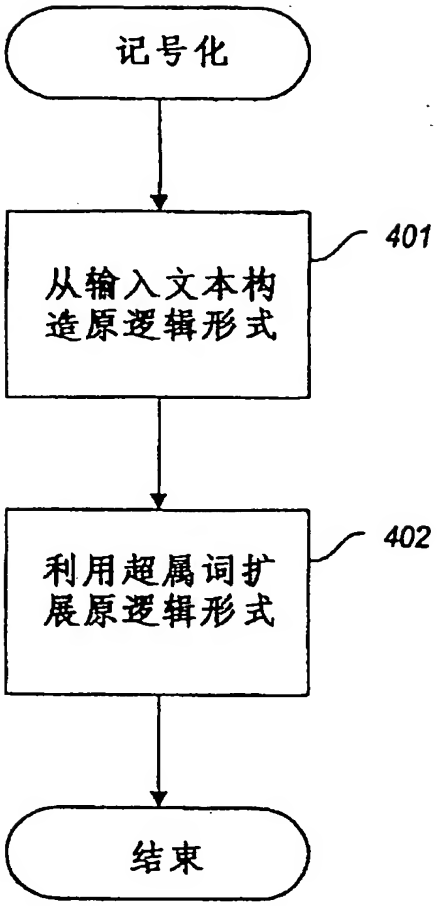


图4

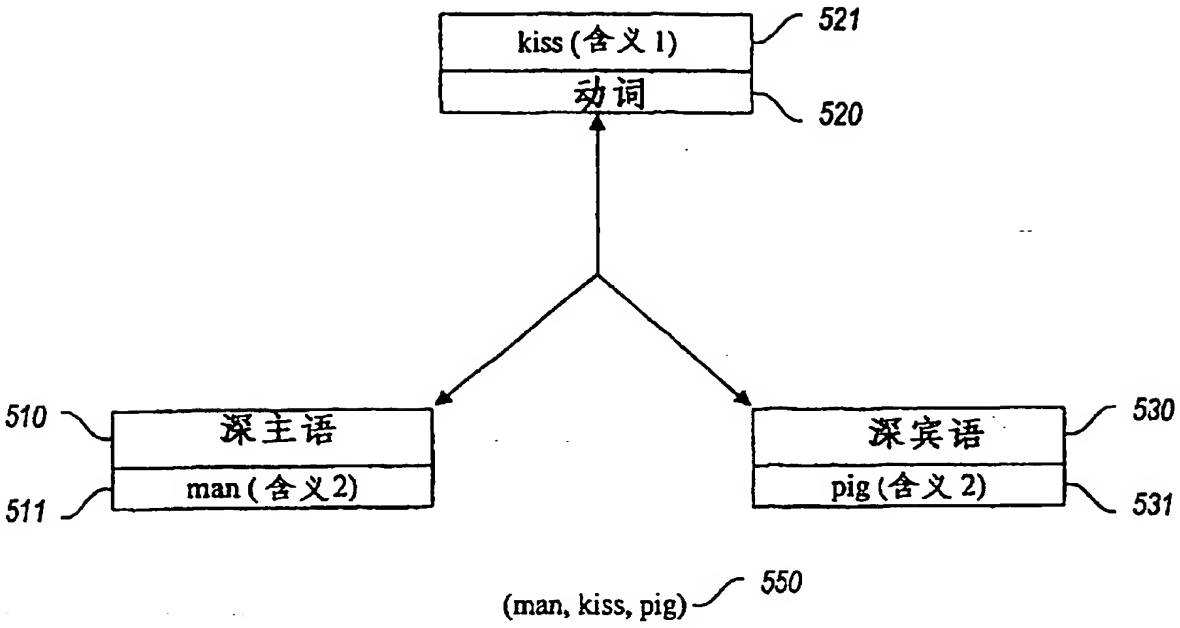

kiss (含义 1)
动词
521
520
510
深主语
man (含义2)
511
深宾语
pig (含义 2)
530
531
(man, kiss, pig)
550

图 5

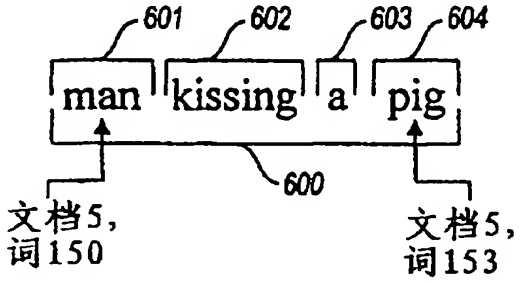

601
602
603
604
man
kissing
a
pig
600
文档5,
词150
文档5,
词153

图 6

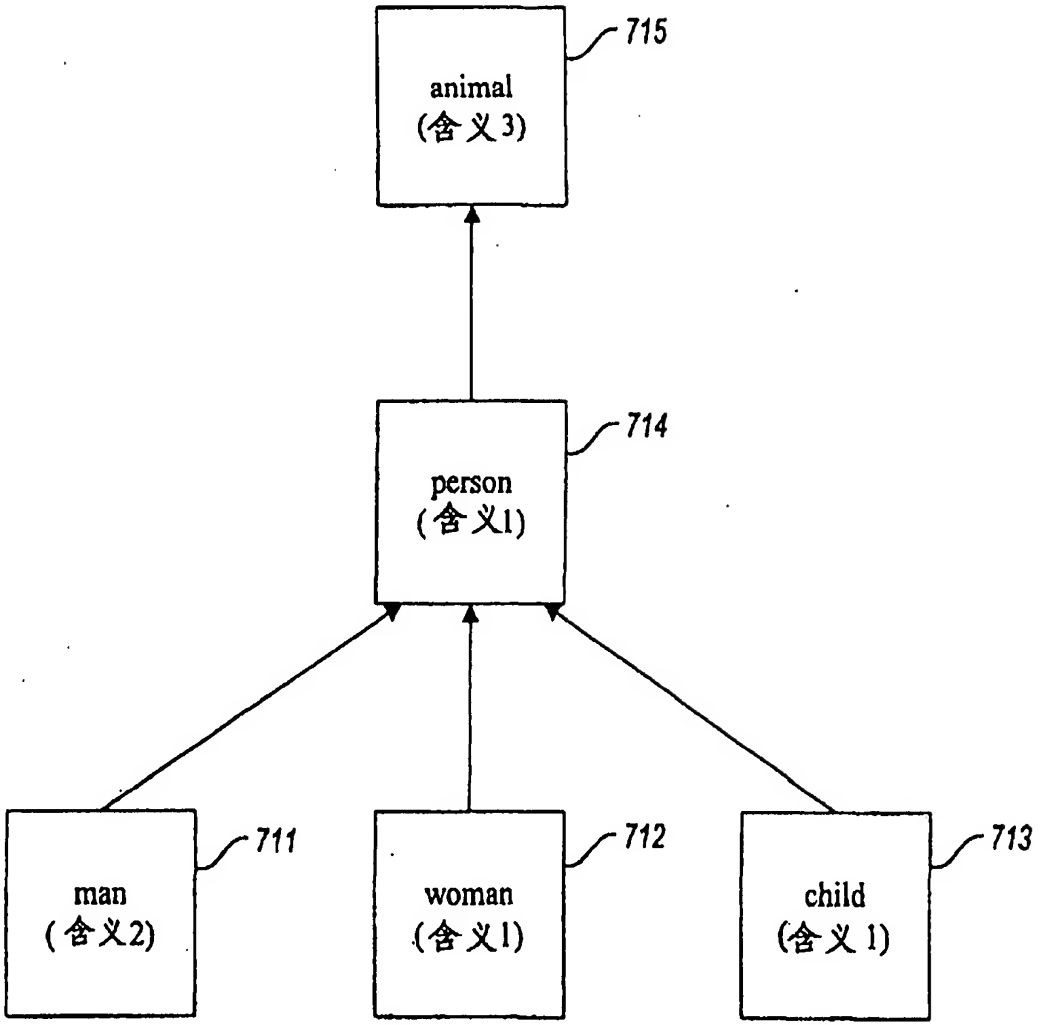

715
animal
(含义3)
714
person
(含义1)
711
man
(含义2)
712
woman
(含义1)
713
child
(含义1)

图 7A

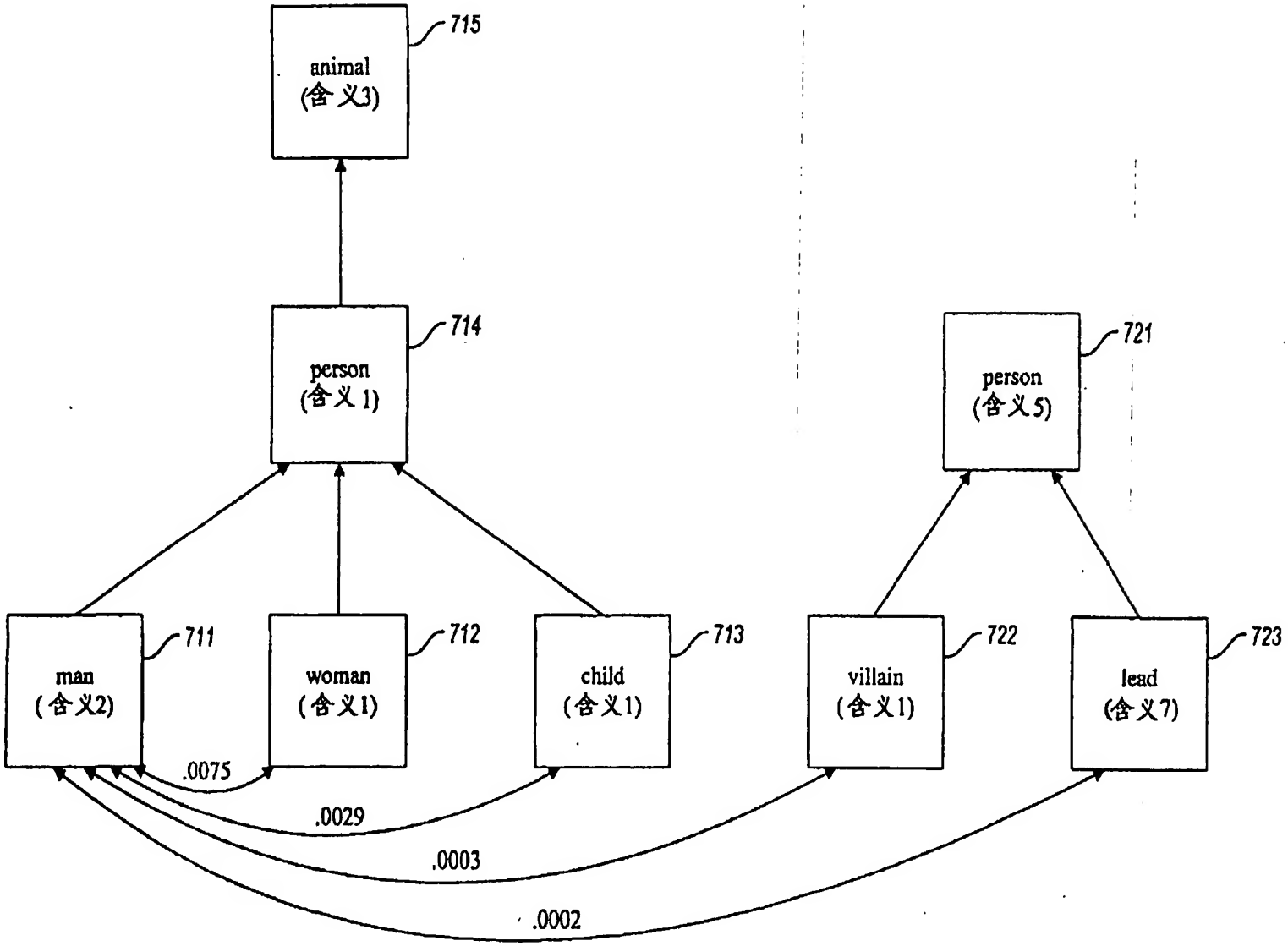

animal
(含义3)
715
person
(含义1)
714
person
(含义5)
721
man
(含义2)
711
woman
(含义1)
712
child
(含义1)
713
villain
(含义1)
722
lead
(含义7)
723
.0075
.0029
.0003
.0002

图7B

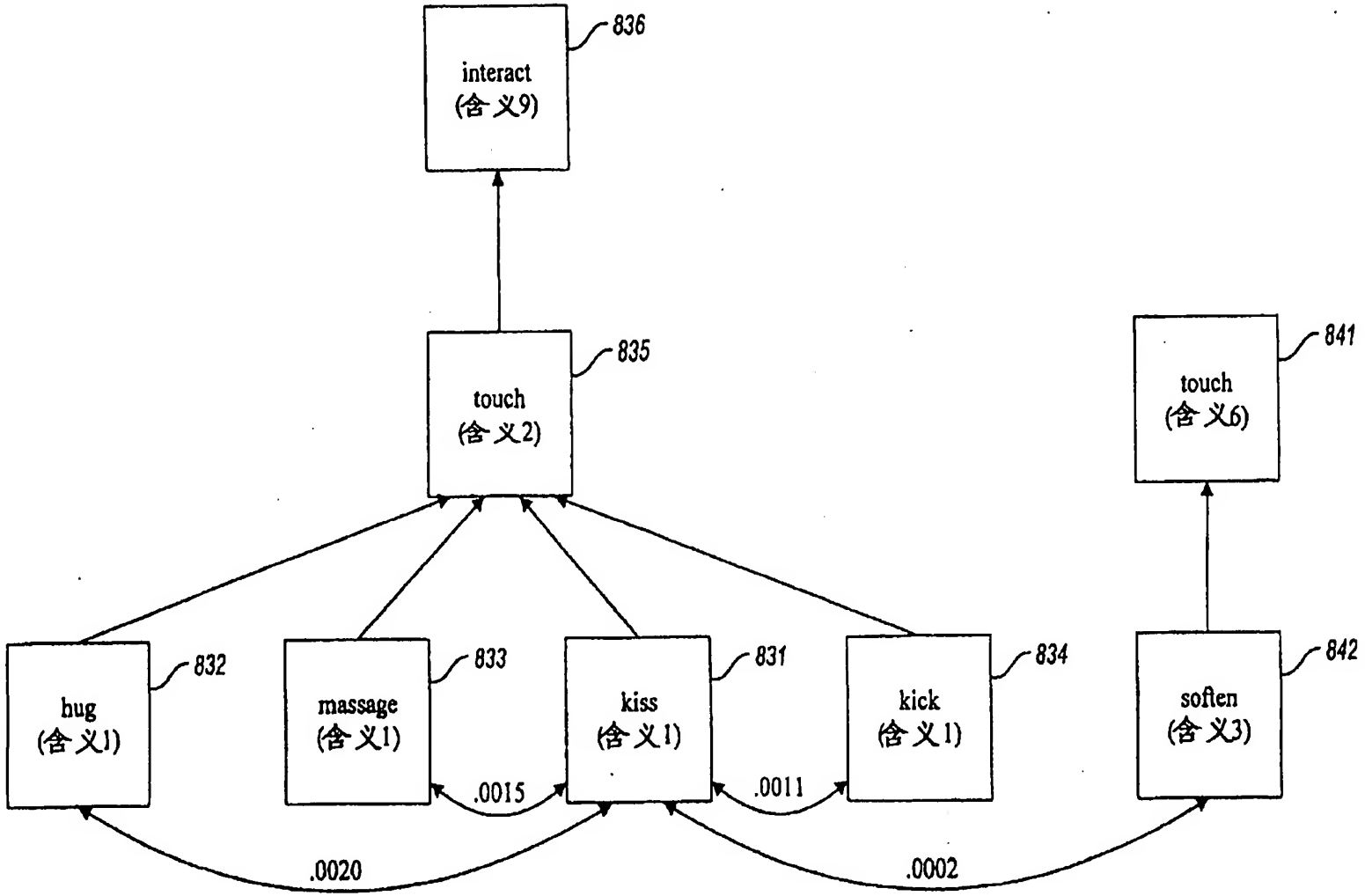

interact
(含义9)
836
touch
(含义2)
835
touch
(含义6)
841
hug
(含义1)
832
massage
(含义1)
833
kiss
(含义1)
831
kick
(含义1)
834
soften
(含义3)
842
.0015
.0011
.0020
.0002

图8

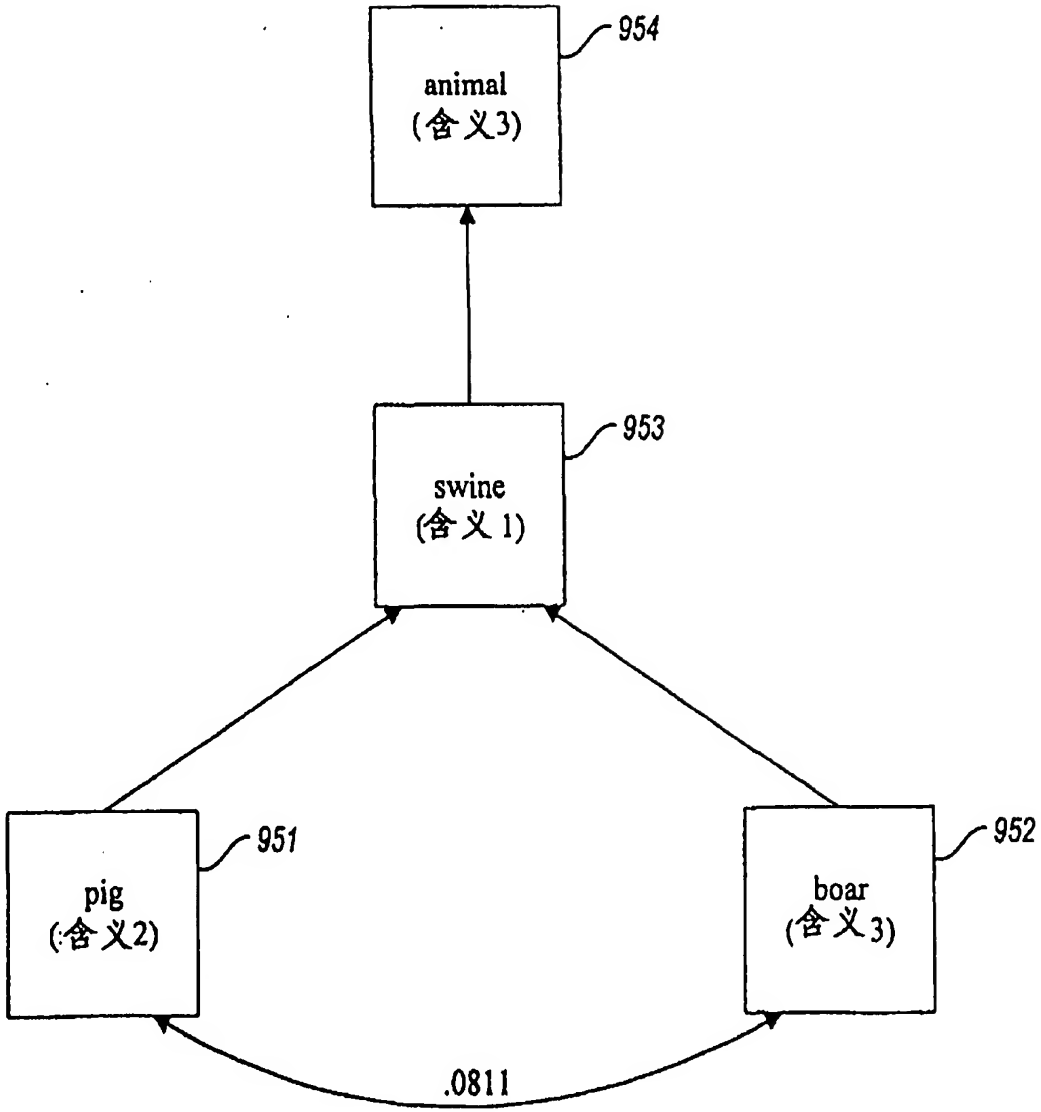


图9

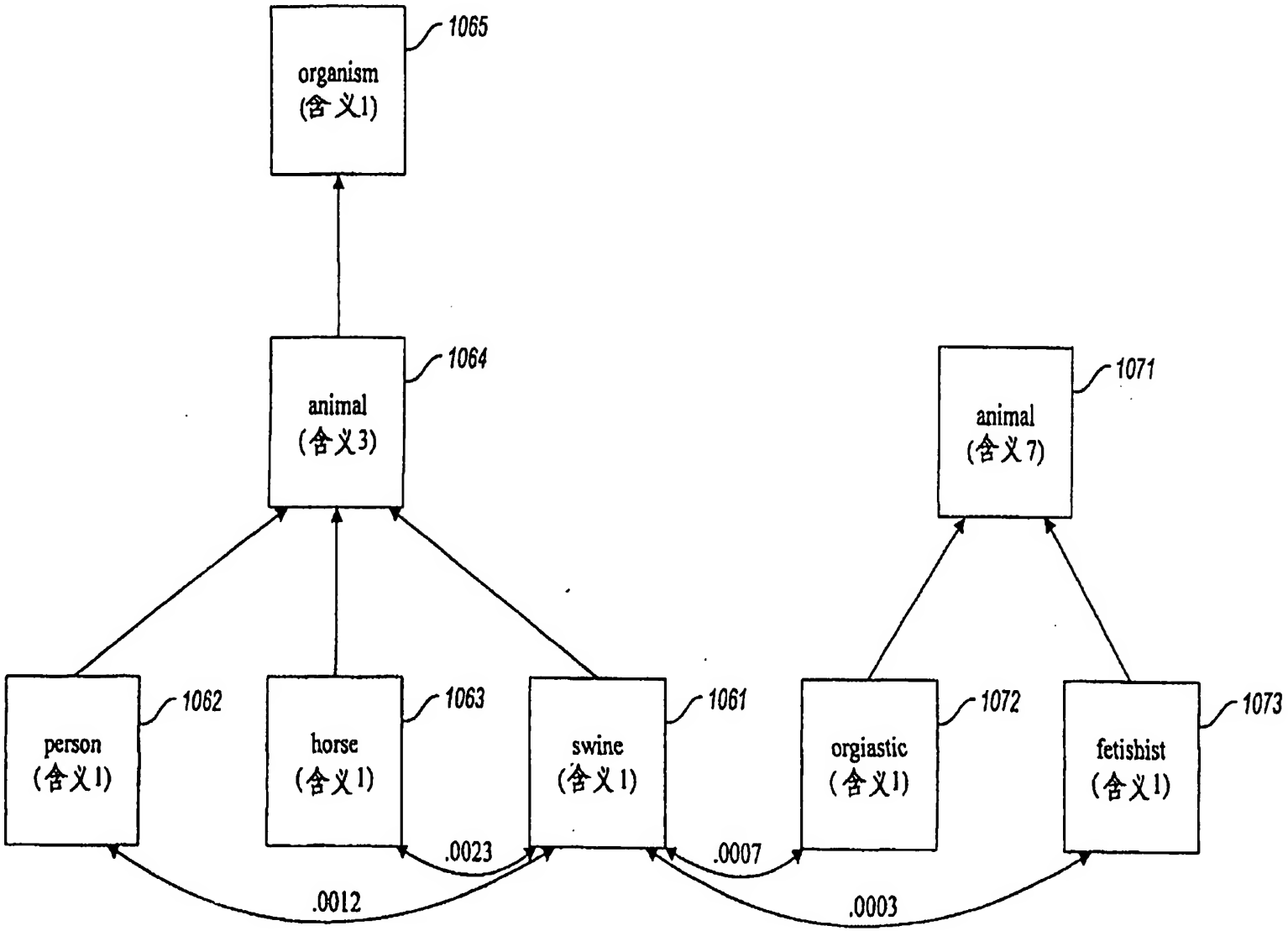

organism
(含义1)
1065
animal
(含义3)
1064
animal
(含义7)
1071
person
(含义1)
1062
horse
(含义1)
1063
swine
(含义1)
1061
orgiastic
(含义1)
1072
fetishist
(含义1)
1073
.0023
.0007
.0012
.0003

图10

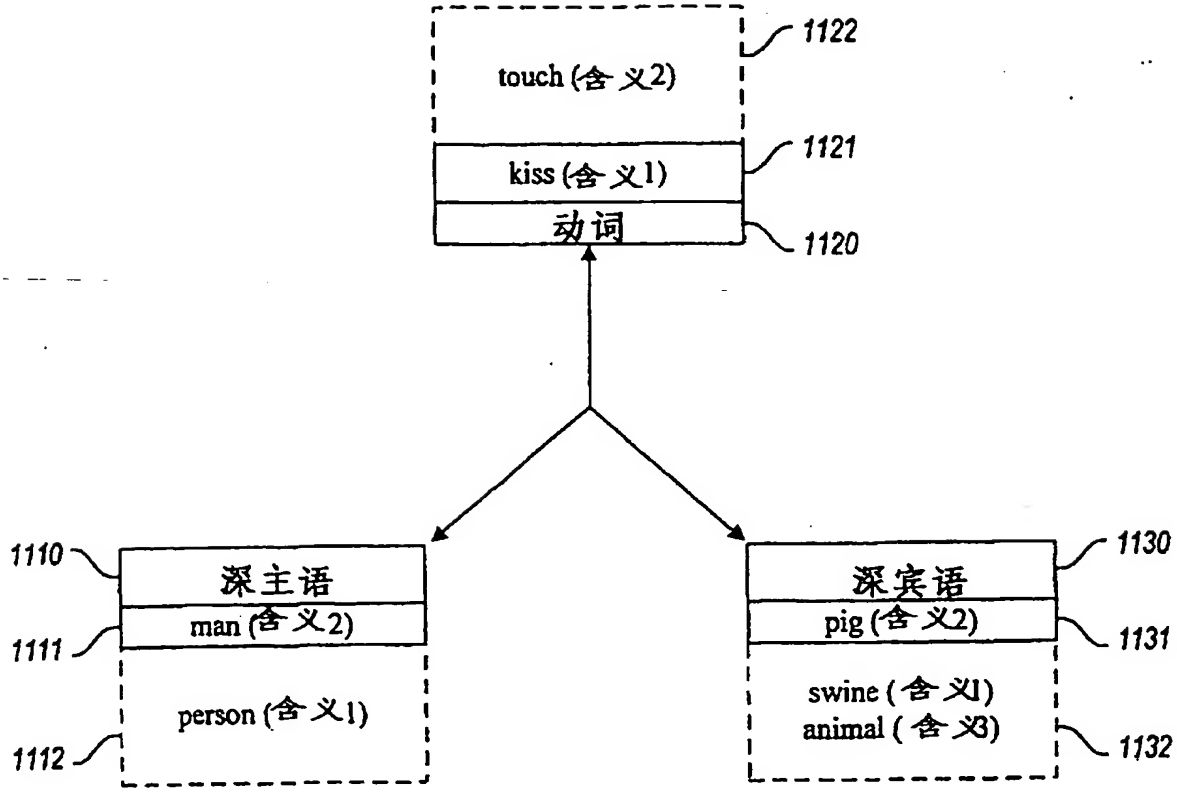

touch (含义2)
1122
kiss (含义1)
1121
动词
1120
1110
深主语
man (含义2)
1111
person (含义1)
1112
深宾语
1130
pig (含义2)
1131
swine (含义1)
animal (含义3)
1132

图11

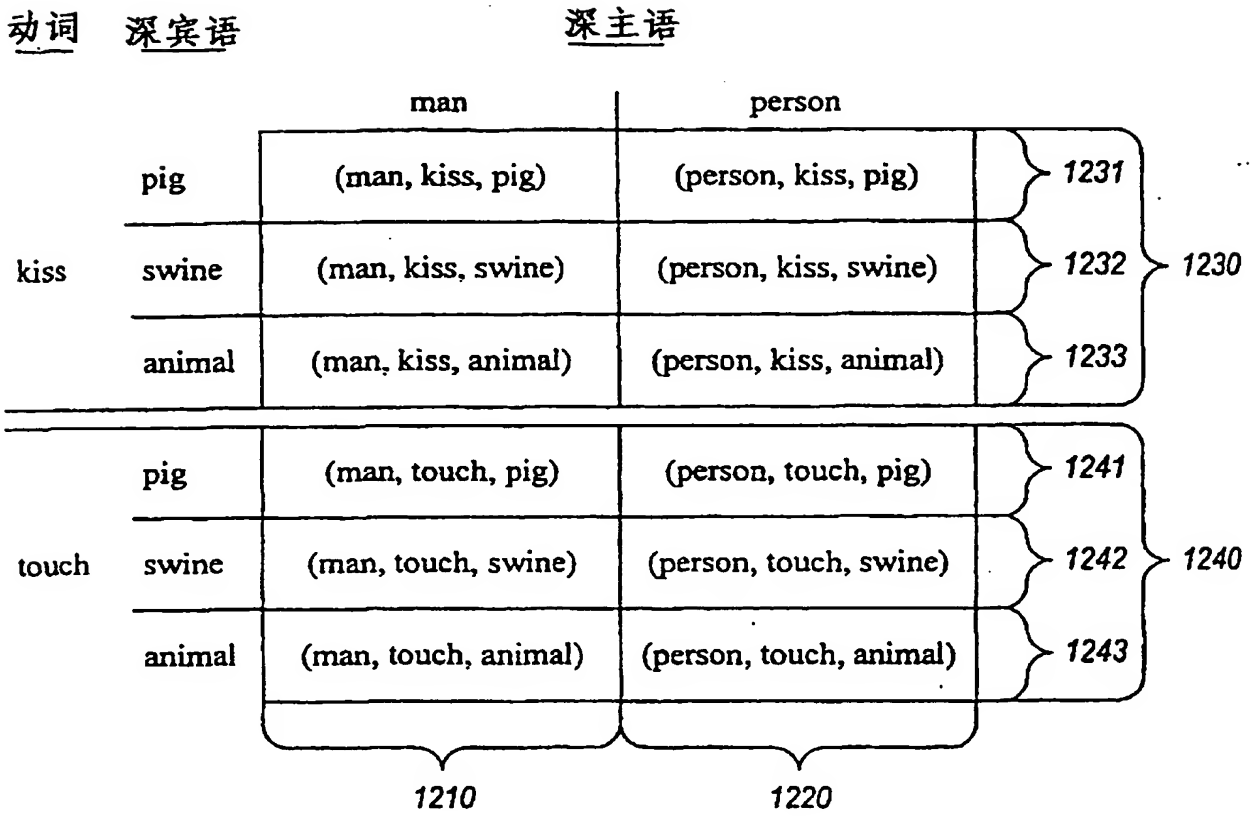

动词
深宾语
深主语
man
person
pig
(man, kiss, pig)
(person, kiss, pig)
1231
kiss
swine
(man, kiss, swine)
(person, kiss, swine)
1232
1230
animal
(man, kiss, animal)
(person, kiss, animal)
1233
pig
(man, touch, pig)
(person, touch, pig)
1241
touch
swine
(man, touch, swine)
(person, touch, swine)
1242
1240
animal
(man, touch, animal)
(person, touch, animal)
1243
1210
1220
1200
((man OR person), (kiss OR touch), (pig OR swine OR animal))

图12

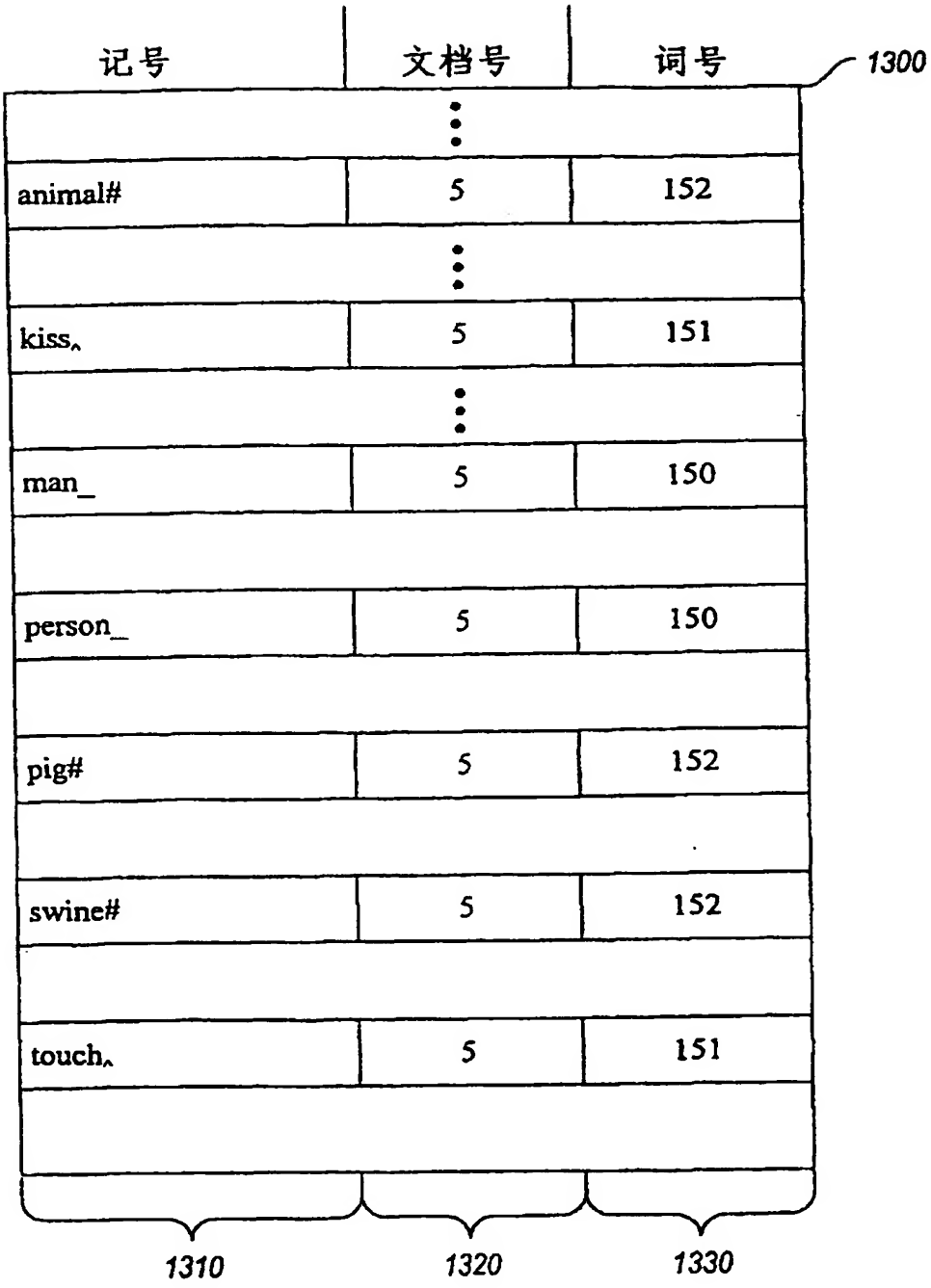

记号
文档号
词号
1300
animal#
5
152
kiss_
5
151
man_
5
150
person_
5
150
pig#
5
152
swine#
5
152
touch_
5
151
1310
1320
1330

图13

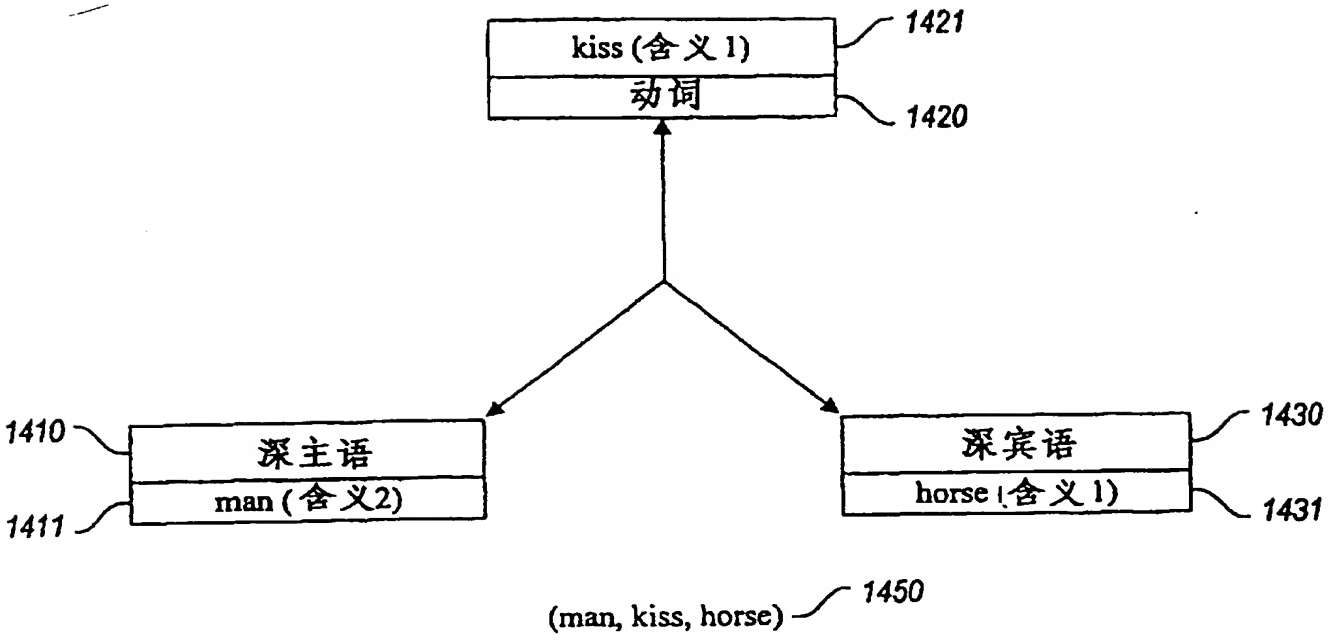

kiss (含义 1)
动词
1421
1420
1410
深主语
man (含义2)
1411
深宾语
horse (含义 1)
1430
1431
(man, kiss, horse)
1450

图14

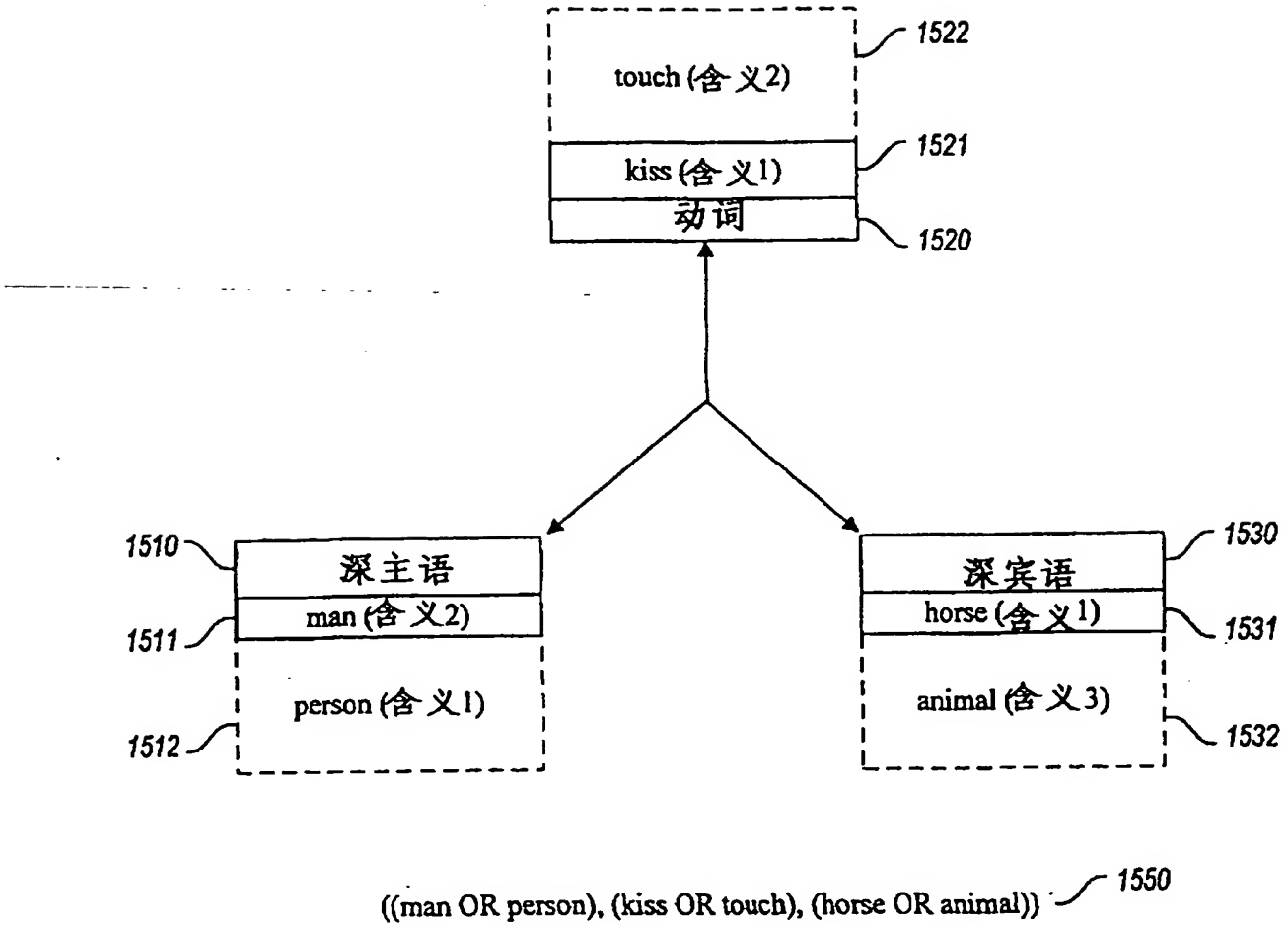

touch (含义2)
1522
kiss (含义1)
1521
动词
1520
1510
深主语
man (含义2)
1511
person (含义1)
1512
深宾语
1530
horse (含义1)
1531
animal (含义3)
1532
((man OR person), (kiss OR touch), (horse OR animal))
1550

图15

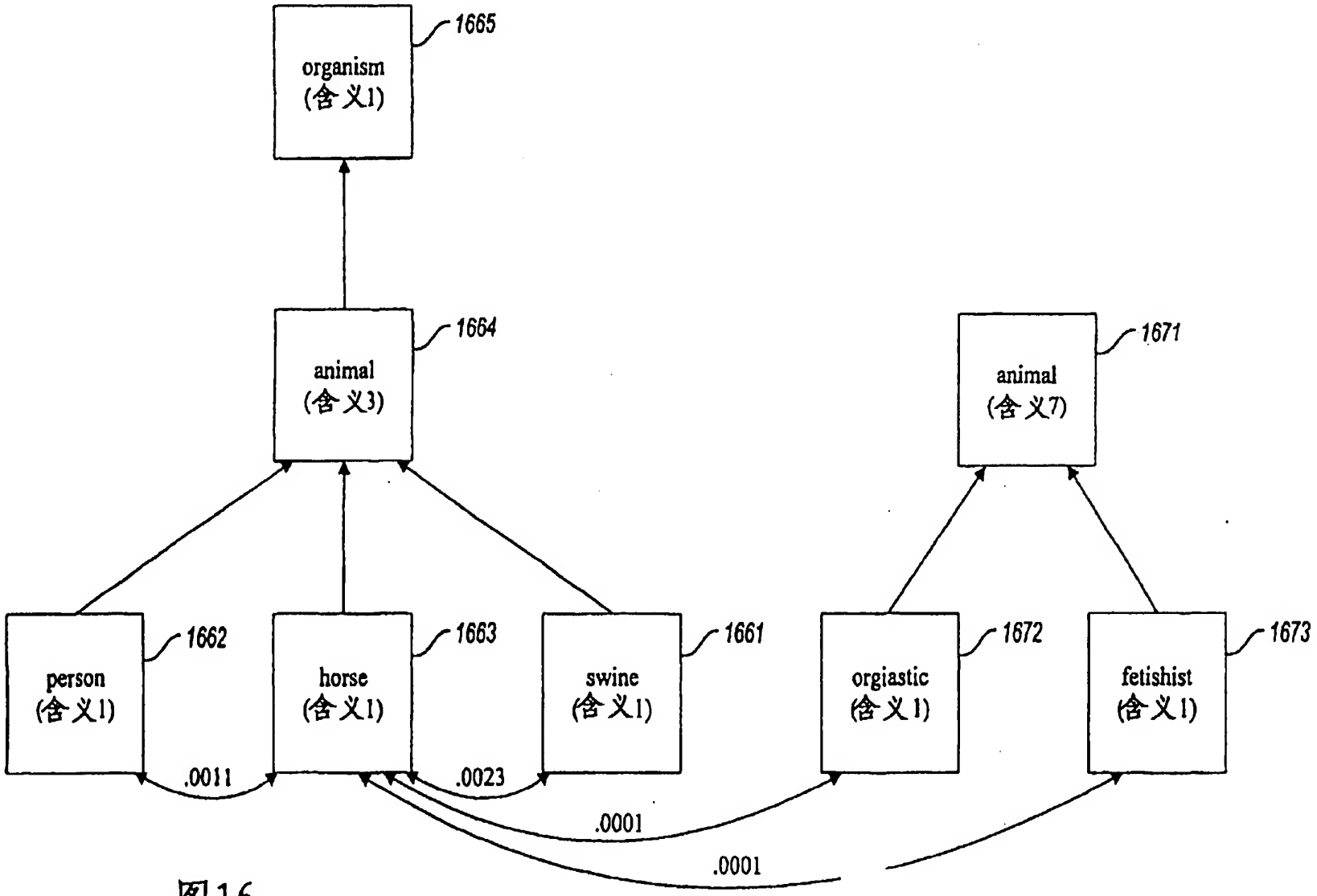

organism
(含义1)
1665
animal
(含义3)
1664
animal
(含义7)
1671
person
(含义1)
1662
horse
(含义1)
1663
swine
(含义1)
1661
orgiastic
(含义1)
1672
fetishist
(含义1)
1673
.0011
.0023
.0001
.0001

图16

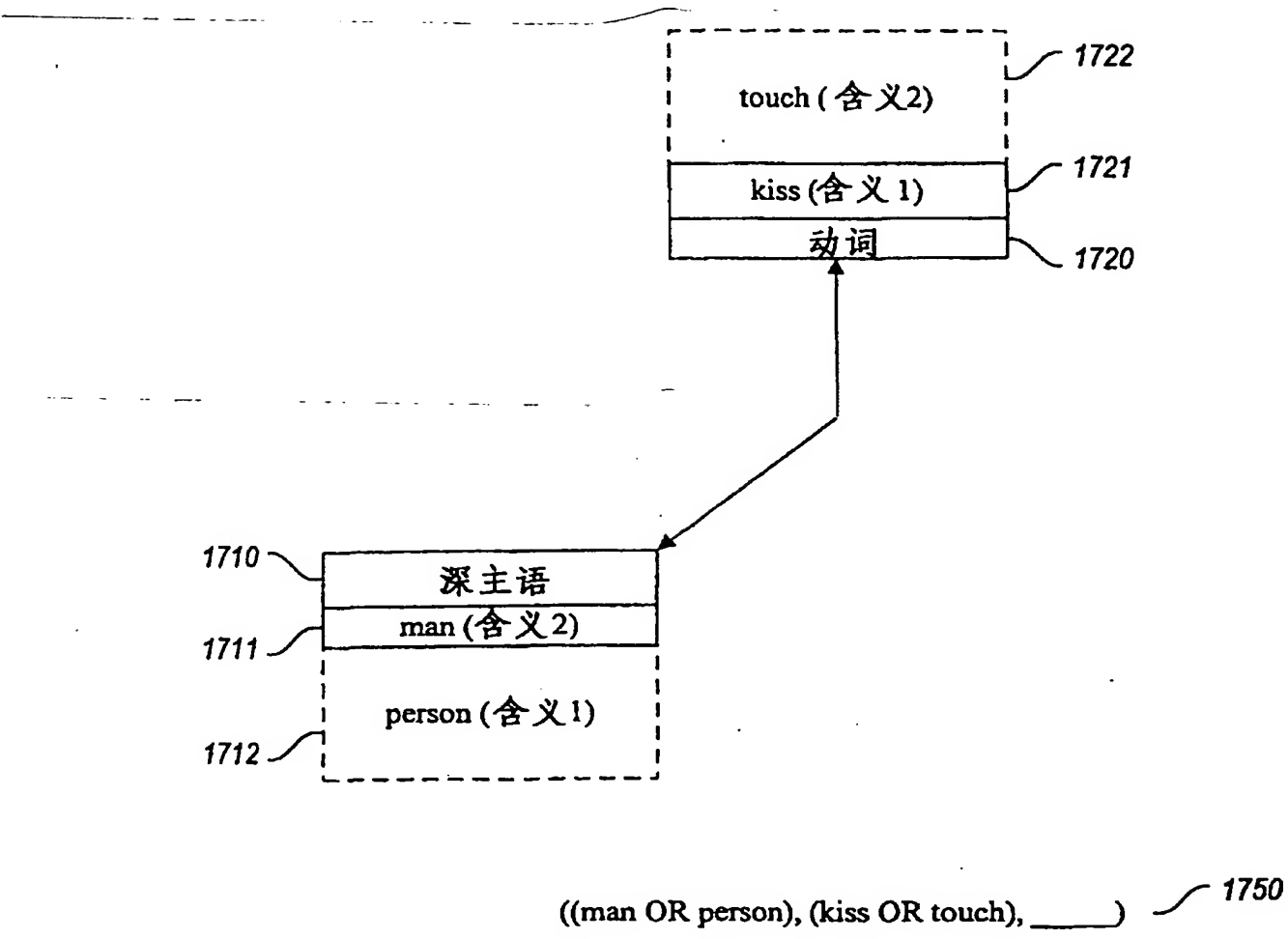

touch (含义2)
1722
kiss (含义1)
1721
动词
1720
1710
深主语
man (含义2)
1711
person (含义1)
1712
((man OR person), (kiss OR touch), _____)
1750

图17

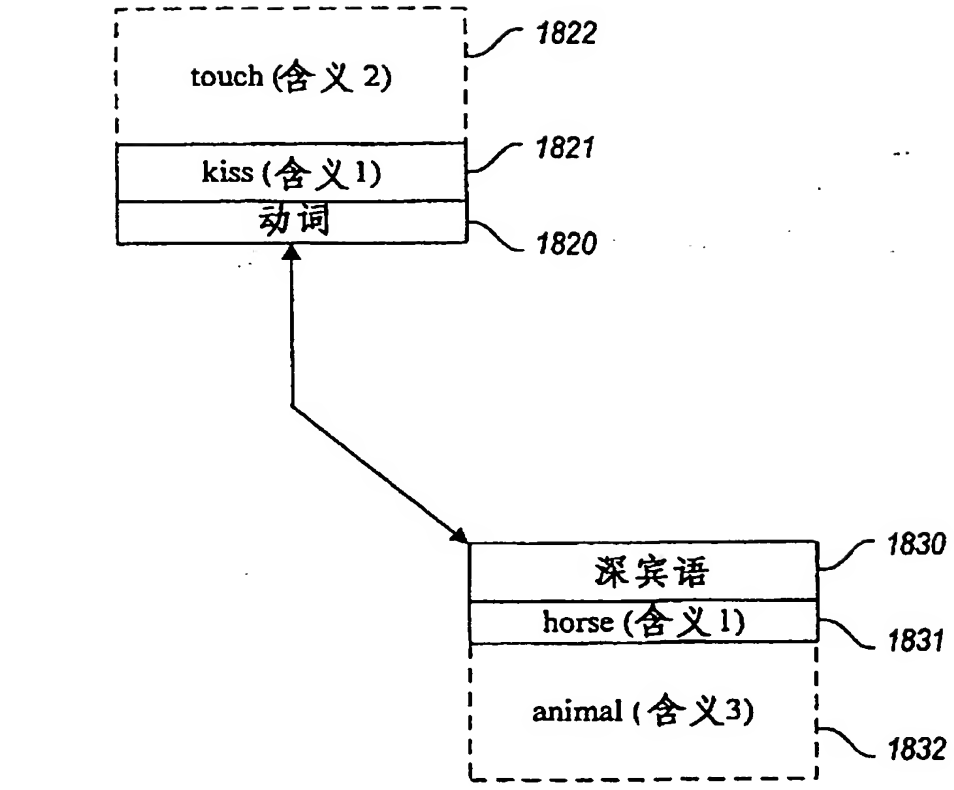

touch (含义 2)
1822
kiss (含义 1)
1821
动词
1820
深宾语
1830
horse (含义 1)
1831
animal (含义3)
1832
(___, (kiss OR touch), (horse OR animal))
1850

图18